In Exploratory Vision: The Active Eye, M. Landy (Ed.), Springer-Verlag, in press.

1

The Synthesis of Vision and Action

Cornelia Fermüller
Yiannis Aloimonos

N00014-93-1-0257

ABSTRACT

Our efforts to reconstruct the world using visual information have led to the insight that the study of Vision should not be separated from the study of a system's actions and purposes. In computational terms this relates to approaching the analysis of perceptual information processing systems through the modelling of the observer and world in a synergistic manner, not through the isolated modelling of observer and world as closed systems. The question still remains: how should such a synergistic modelling be realized? This chapter addresses the question by providing a methodology for synthesizing vision systems and integrating perception and action. In particular, we outline an architecture for purposive vision systems and present a hierarchy of navigational competences based on computational models of increasing complexity, employing representations of motion, shape, form and space. Pure computational considerations will not tell us what visual competences and representations are important to vision systems performing a set of tasks. Interaction, however, with empirical sciences such as Neurobiology, Physiology, Psychology, Ethology, etc., can give us inspiration about the visual categories relevant to systems existing in real world environments. Throughout the chapter, we describe biological findings and how they affect the choice of computational models and representations needed for the synthesis of a hierarchy of navigational competences in a working system.

1.1 Prolegomena

A complete theory of perception must examine a broad set of topics and their interaction, ranging from the environment and the stimuli to sensory organs, the brain and the tasks supported by perception. Many theories for explaining visual perception have been proposed over the centuries. For

[0]Computer Vision Laboratory, Center for Automation Research, Department of Computer Science, and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742-3275

tractability reasons, these theories concentrated on only one or a few of the topics mentioned above. They range from the theory of Empedocles (440 B.C.) to the computational theory of David Marr (1982); some of the them are well known to students of perception (Gestaltist theories (Kohler, 1947), the theory of the empiricist Helmholtz (1896), Gibson's theory of direct perception (1979)), while others are almost forgotten (e.g., Brunswick's theory of probabilistic functionalism (1956)).

The development of each theory, as is the case in all disciplines, was influenced by previous ones and other philosophical or scientific ideas prominent at the time. For example, the Gestalt theories were influenced by Kant's (1990) Critique of Pure Reason (although 100 years after its publication) and Marr's computational theory by neuroanatomical developments of his time (Hubel & Wiesel, 1968) as well as by initial results on visual agnosia (Warrington & Shallice, 1984). In our days, more than ever before, our views on the architecture and the structure of the computational mechanisms underlying the behavior of intelligent organisms or robots possessing perception, are influenced by advances in various disciplines that study the brain.

During the 1960's and 70's it was commonly supposed that each visual area of the cerebral cortex analyzes all the information in the field of view, but at a more complex level than the antecedent area, the areas forming a sort of hierarchical chain . The computational theories that appeared at these times reflected this doctrine and efforts were made for discovering mechanisms to reconstruct general descriptions of the visible scene (Marr, 1982; Horn, 1986; Aloimonos & Shulman, 1993). Vision was studied in a disembodied manner by concentrating mostly on *stimuli*, *sensory organs* and the *brain*. In some sense, this work argued that "seeing" and "thinking" are distinguishable activities, with "seeing" being a mechanical act that does not originate anything (Nalwa, 1993; Kanizsa, 1979). This in turn contributed to the separation of main stream Artificial Intelligence (that was about "thinking") and Computer Vision (that was about "seeing").

During the 1980's with the emergence of Active Vision (Bajcsy, 1988; Aloimonos, 1993; Aloimonos, Weiss & Bandopadhay, 1988; Ballard, 1991), researchers considered, in addition to the topics of stimuli, sensors and brain, the topics of reflexes and motor responses and in particular the motor responses whose goal is to control the image acquisition process. The realization that active vision systems could, with selective perception, perform a number of interesting tasks contributed to a novel view of the global structure of a "seeing" system. Several researchers today view a vision system as consisting of a set of behaviors, programs capable of supporting a set of actions. It has been understood that Perception should not be studied in isolation, but in conjunction with the physiology of systems and with the tasks that systems perform. In the discipline of Computer Vision such ideas caused researchers to extend the scope of their field. If before Computer Vision was limited to the study of mappings of a given set of vi-

sual data into representations on a more abstract level, it now has become clear that Image Understanding should also include the process of selective acquisition of data in space and time. This has led to a series of studies published under the headings of Active, Animate, Purposive, or Behavioral Vision. A good theory of vision would be one that can create an interface between perception and other cognitive abilities. However, with a formal theory integrating perception and action still lacking, most studies treated Active Vision (Aloimonos, Weiss & Bandopadhay, 1988; Bajcsy, 1988) as an extension of the classical reconstruction theory, employing activities only as a means to regularize the ill-posed classical inverse problems.

1.2 Marr's theory and its drawbacks

Let us summarize the key features of the classical theory of Vision in order to point out its drawbacks as an overall framework for studying and building perceptual systems: In the theory of Marr (1982), the most influential in recent times, Vision is described as a reconstruction process, that is, a problem of creating representations of increasingly high levels of abstraction, leading from 2-D images over the primal sketch through the $2\frac{1}{2}$-D sketch to object centered descriptions ("from pixels to predicates") (Pentland, 1986). Marr suggested that visual processes—or any perceptual/cognitive processes—are information processing tasks and thus should be analyzed at three levels: (a) at the computational theoretic level (definition of the problem and its boundary conditions; formulation of theoretical access to the problem), (b) at the level of selection of algorithms and representations (specification of formal procedures for obtaining the solution), and (c) at the implementational level (depending on the available hardware).

In the definition of cognitive processing in the classical theory, Vision is formalized as a pure information processing task. Such a formalization requires a well-defined closed system. Since part of this system is the environment, the system would be closed only if it were possible to model all aspects of objective reality. The consequence is well-known: Only toy problems (blocks worlds, Lambertian surfaces, smooth contours, controlled illumination, and the like) could be successfully solved.

The strict formalization of representations at different levels of abstraction gave rise to breaking the problems into autonomous subproblems and solving them independently. The conversion of external data (sensor data, actuator commands, decision making, etc.) into an internal representation was separated from the phase of algorithms to perform computations on internal data; signal processing was separated from symbolic processing and action. Processing of visual data was treated, for the most part, in a syntactic manner and semantics was treated in a purely symbolic way using the results of the syntax analysis. This is not surprising, since Computer Vision was considered as a subfield of Artificial Intelligence and thus

studied using the same methodology, influenced by the ideas about computational theories of the last decades (Ernst & Newell, 1969; Gelernter, 1959; Nilsson, 1980).

The strict hierarchical organization of representational steps in the Marr paradigm makes the development of learning, adaptation and generalization processes practically impossible (no doubt there hasn't been much work on computational "vision and learning"). Furthermore, the conceptualization of a vision system as consisting of a set of modules recovering general scene descriptions in a hierarchical manner introduces computational difficulties with regard to issues of robustness, stability, and efficiency. These problems lead us to believe that general vision does not seem to be feasible. Any system has a specific relationship with the world in which it lives, and the system itself is nothing but an embodiment of this relationship. In the Marr approach the algorithmic level has been separated from the physiology of the system (the hardware) and thus vision was studied in a disembodied transcendental manner.

Of course, many of the solutions developed for disembodied systems may also be of use for embodied ones. In general, however, this does not hold. Having available an infinite amount of resources, every (decidable) problem can be solved in principle. Assuming that we live in a finite world and that we have a finite number of possibilities for performing computations, any vision problem might be formulated as a simple search problem in very high dimensional space. From this point of view, the study of embodied systems is concerned with the study of techniques to make seemingly intractable problems tractable.

Not the isolated modelling of observer and world (as closed systems) but the modelling of observer and world in a synergistic manner, will contribute to the understanding of perceptual information processing systems (Sommer, 1994). The question, of course, still remains how such a synergistic modelling should be realized. Or: How can we relate perception and action? What are the building blocks of an intelligent perceptual system? What are the categories into which the system divides its perceptual world? What are the representations it employs? How is it possible to implement such systems in a flexible manner to allow them to learn from experience and extend themselves to better ones? In this paper we present a formal framework for addressing these questions. Our exposition describes both recent technical results and some of our future research agenda.

1.3 The Architecture

1.3.1 The modules of the system

At the highest level of abstraction, a vision system consists of a set of maps (Fig. 1.1) that relate the observer's space-time representations. For

the purpose of a more systematic study, we classify the maps into three categories: the visual competences, the action routines, and the learning procedures. We distinguish two kinds of representations according to the amount of processing performed on them: representations of the perceptual information computed from visual input, and representations of any kind of perceptual information acquired over time and shared and organized in memory. Fig. 1.2 gives a more detailed description of a purposive vision system.

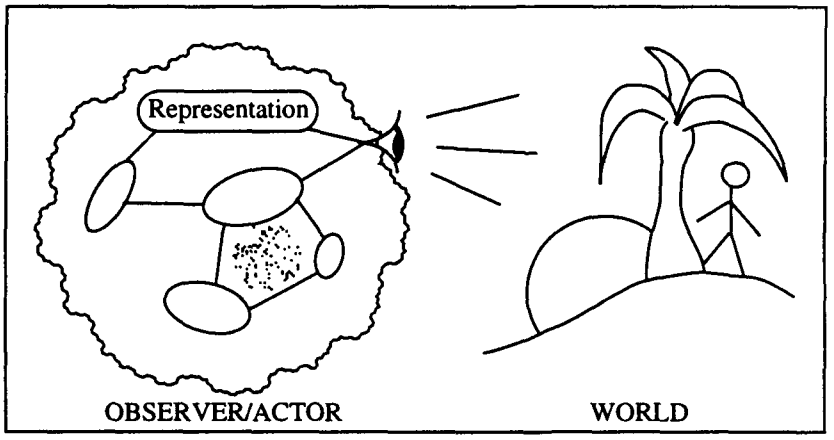

FIGURE 1.1. An intelligent system with vision creates various representations of space-time that it uses in order to perform various actions.

At a different level of abstraction that modularizes the system, Fig. 1.3 describes the basic components of a purposive vision system. The abstract procedures and representations of a vision system are: the procedures for performing visual perceptions, physical actions, learning, and information retrieval, and purposive representations of the perceptual information along with representations of information acquired over time and stored in memory.

At any time a purposive vision system has a goal or a set of goals it approaches as best as it can by means of its available resources. Thus at any time the system is engaged in executing a task. The visual system possesses a set of visual competences with which it processes the visual information. The competences compute purposive representations. Each of these representations captures some aspect of the total visual information. Thus compared with the representations of the old paradigm, they are partial. The representations are of different complexities with regard to the space they describe. The purposive representations themselves are purposive descriptions of the visual information organized in certain data structures. The purposive representations access programs which we call "action routines." This collective name refers to two kinds of routines; the first kind are

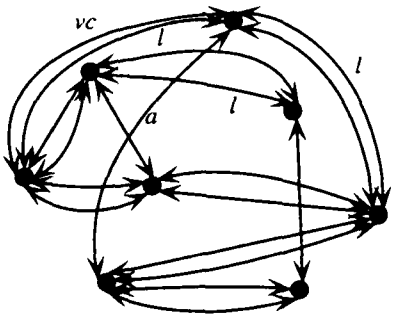

FIGURE 1.2. An intelligent system with vision consists of a map of maps. These maps map different representations of space-time (•) into each other. Space-time includes, of course, the system itself. In order to study these maps more systematically, we divide them into three categories: the visual competences (vc), the action routines (a), and the learning procedures (l).

the programs that schedule the physical actions to be performed, i.e., they initialize motor commands and thus provide the interface to the body, and the second kind schedule the selection of information to be retrieved from the purposive representations and stored in long-term memory. An important aspect of the architecture is that the access of the visual processes to the actions is on the basis of the contents of the purposive representations, i.e., the contents of the purposive representations serve as addresses to the actions. Another class of programs is responsible for learning by providing the actions, the competences, and the representations with the means to change and adjust parameters.

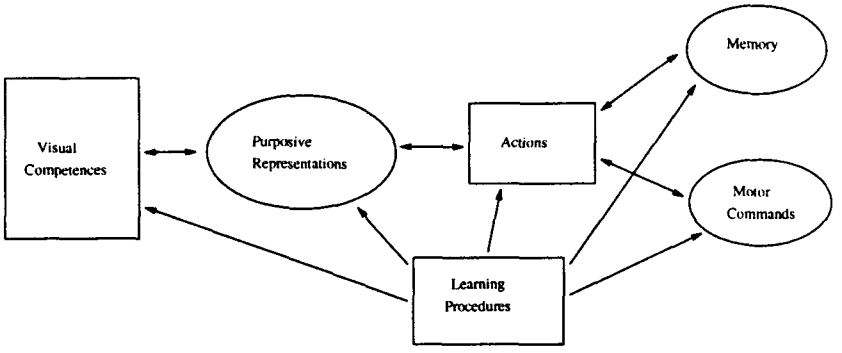

FIGURE 1.3. Working model: Basic components of a purposive vision system.

As can be seen from the figure, learning takes place at various levels of, as well as in between, the modules of the system. For a flexible vision system, it should be possible to learn the parameters describing actions, to acquire new actions, to learn parameters describing visual competences, to acquire new visual competences that compute new purposive representations, and to learn the sequences of actions and perceptual competences to perform a task. In any case, learning is accomplished by means of programs—learning procedures—that allow the change and adaptation of parameters in order to learn competences, actions, and their interrelationships.

The purposive perceptual representations, as well as representations containing other kinds of information, are stored in memory. The storing must happen in an efficient way according to the available memory space. Different representations share common elements. Memory organization techniques that allow you to store information according to its content are needed. Also, designing a memory for representations includes designing the procedures necessary for fast and reliable access. In this chapter we focus our discussion on the visual competences.

Let us summarize in which way the above model captures the study of perception and action in a synergistic way, and address some of the questions posed in Section 1.2: In this model the intelligence of a purposive system is embodied in its visual competences and its actions. Thus competences and actions are considered to be the building blocks of an intelligent system. In order to meet a purpose (a task which is stated in the form of events that can be perceived by means of the perceptual processes), a system executes behaviors. Thus, behaviors, which are an emergent attribute of the system, couple perception and action. They constitute some form of structural adaptation which might either be visible externally or take place only internally in the form of parameter adaptation.

1.3.2 Outline of the approach

If we aim to understand perception, we have to come up with some methodology to study it. The ideal would be to design a clearly defined model for the architecture of vision systems and start working on its components. However, we have few answers available when it comes to actually talking about the visual categories that are relevant for visual systems. The kind of representations needed to perform a task depends on the embodiment of the system and the environment in which it lives. Answers to these questions cannot come only as insight gained from the study of mathematical models. There must be empirical studies investigating systems (biological and artificial ones) that will tell us how to couple functionality, visual categories and visual processes. If we haven't understood how we actually could develop visual competences for systems that work in environments as complex as our own, we won't be able to obtain a global view of the overall architecture and functionality of vision systems. At this time it also

wouldn't contribute much to the development of our understanding to just develop particular systems that perform particular tasks, for example, a system that recognizes tables. Even if we were able to create such a system having a success rate of 99%, it would have the capacity of recognizing many things that are unknown to us, and not just tables. Thus, by aiming to build systems that recognize certain categories that seem relevant to our symbolic language repertoire, we wouldn't gain much insight into perception.

It seems somehow natural that the only way out of this problem of where to start is to approach the study of vision systems in an "evolutionary" way. We call such an approach the synthetic (evolutionary) approach. We give here a short outline of the ideas behind this approach, which we discuss in detail in the remainder of the paper. It means we should start by developing individual primitive visual operations and provide the system in this way with visual capabilities (or competences). As we go on, the competences will become more and more complex. At the same time, as soon as we have developed a small number of competences, we should work on their integration. Such an endeavor throws us immediately into the study of two other major components of the system. How is visual information related to action and how is the information represented? How is it organized and how it is coordinated with the object recognition space? We are confronted on the one hand with the study of activities and the integration of vision and action, and on the other hand with the study of the memory space with all its associated problems of memory organization, visual data representation, and indexing—the problem of associating data stored in the memory with new visual information. Furthermore, we also have to consider the problem of learning from the very beginning.

1.4 The competences

1.4.1 Computational Principles

A: Model-Complexity

Our goal is to analyze, in order to design, a system from a computational point of view. We argued earlier that the study of visual systems should be performed in a hierarchical manner according to the complexity of the visual processes. As a basis for its computations a system has to utilize mathematical models, which serve as abstractions of the representations employed. Thus, when referring to the complexity of visual processes, we mean the complexity of the mathematical models involved.

The synthetic approach calls first for studying capabilities whose development relies on only simple models and then going on to study capabilities requiring more complex models. Simple models do not refer to environment-

or situation-specific models which are of use in only a limited number of situations. Each of the capabilities requiring a specified set of models should be used for solving a well-defined class of tasks in every environment and situation to which the system is exposed. If our goal is to pursue the study of perception in a scientific way, as opposed to industrial development, we have to accept this requirement as one of the postulates, although it is hard to achieve. Whenever we perform computations, we design models on the basis of assumptions, which in the case of visual processing are constraints on the space-time in which the system is acting, on the system itself, and on their relationship. An assumption, however, can be general with regard to the environment and situation, or very specific.

For example, the assumption about piecewise planarity of the world is general with regard to the environment (every continuous differentiable function can be approximated in an infinitesimal area by its derivatives). However, in order to use this assumption for visual recovery, additional assumptions regarding the number of planar patches have to be made; these are environment-specific assumptions. Similarly, we may assume that the world is smooth between discontinuities; this is general with regard to the environment. Again, for this assumption to be utilized we must make some assumptions specifying the discontinuities, and then we become specific. We may assume that an observer only translates. If indeed the physiology of the observer allows only translation, then we have made a general assumption with regard to the system. If we assume that the motion of an observer in a long sequence of frames is the same between any two consecutive frames, we have made a specific assumption with regard to the system. If we assume that the noise in our system is Gaussian or uniform, again we have made a system-specific assumption.

Our approach requires that the assumptions used be general in regard to the environment and the system. Scaled up to more complicated systems existing in various environments, this requirement translates to the system's capability to decide whether a model is appropriate for the environment in which the system is acting. A system might possess a set of processes that together supply the system with one competence. Some of the processes are limited to certain environmental specifications. Thus, the system must be given the capability to acquire knowledge about what processes to apply in a specific situation.

The motivation for studying competences in a hierarchical way is to gain increasing insight into the process of vision, which is extremely complex. Therefore the capabilities which require more complex models should be based on "simpler," already developed capabilities. The complexity of a capability is given by the complexity of the assumptions employed; what has been considered a "simple" capability might require complex models, and vice versa.

B: Qualitative models

The basic principle concerning the implementation of processes subserving the capabilities, which is motivated by the need for robustness, is the quest for algorithms which are qualitative in nature. We argue that visual competences should not be formulated as processes that reconstruct the world but as recognition procedures. Visual competences are procedures that recognize aspects of objective reality which are necessary to perform a set of tasks. The function of every module in the system should constitute an act of recognizing specific situations by means of primitives which are applicable in general environments. Each such entity recognized constitutes a category relevant to the system. Following are some examples from navigation.

The problem of independent motion detection by a moving observer usually has been addressed with techniques for segmenting optical flow fields. But it also may be tackled through the recognition of non-rigid flow fields for a moving observer partially knowing its motion (Aloimonos, 1990; Nelson, 1991; Thompson & Pong, 1990). Pursuing a target amounts to recognizing the target's location on the image plane along with a set of labels representing aspects of its relative motion sufficient for the observer to plan its actions. Motion measurements of this kind could be relative changes in the motion such as a turn to the left, right, above, or down; or focusing further away, or closer. In the same way, the problem of hand/eye coordination can be dealt with using stereo and other techniques to compute the depth map and then solve the inverse kinematics problem in order to move the arm. While the arm is moving the system is blind (Brady, Hollerbach, Johnson, Lozano-Perez & Mason, 1983); however, the same problem can be solved by creating a mapping (the perceptual kinematic map) from image features to the robot's joints. The positioning of the arm is achieved by recognizing the image features (Hervé, 1993).

Instead of reconstructing the world, the problems described above are solved through the recognition of entities that are directly relevant to the task at hand. These entities are represented by only those parameters sufficient to solve the specific task. In many cases, there exists an appropriate representation of the space-time information that allows us to derive directly the necessary parameters by recognizing a set of locations on this representation along with a a set of attributes. Since recognition amounts to comparing the information under consideration with prestored representations, the described approaches to solving these problems amount to matching patterns.

In addition, image information should be utilized globally whenever possible. Since the developed competences are meant to operate in real environments the computations have to be insensitive to errors in the input measurements. This postulates a requirement for redundancy in the input used. The partial information about the scene, which we want to recognize,

mostly will be globally encoded in the image information. The computational models we are using should be such that they map global image information into partial scene information. Later in this section we will demonstrate our point by means of the rigid motion model.

In order to speak of an algorithm as qualitative, the primitives to be computed do not have to rely on explicit unstable, quantitative models. Qualitativeness can be achieved in a number of ways: The primitives might be expressible in qualitative terms, or their computation might be derived from inexact measurements and pattern recognition techniques, or the computational model itself might be proved stable and robust in all possible cases.

The synthetic approach has some similarities at the philosophical level with Brooks' proposal for understanding intelligent behavior through the construction of working mechanisms (1986). In proposing the subsumption architecture, Brooks suggested a hierarchy of competences such as avoiding contact with objects, exploring the world by seeing places, reasoning about the world in terms of identifiable objects, etc. This proposal, however, did not provide a systematic way of creating a hierarchy of competences by taking into account the system's purpose and physiology. This is the relevant question.

1.4.2 Biological hierarchy

It remains to be discussed what these simple capabilities actually are on which we should concentrate our first efforts. Other scientific disciplines give us some answers. Much simpler than the human visual system are the perceptual systems of lower animals like medusae, worms, crustaceans, insects, spiders and molluscs. Researchers in neuroethology study such systems and by now have gained some understanding of them. Horridge (1987, 1991), working on insect vision, studied the evolution of visual mechanisms and proposed hierarchical classifications of visual capabilities. He argued that the most basic capabilities found in animals are based on motion. Animals up to the complexity of insects perceive objects entirely by relative motion. His view concerning the evolution of vision is that objects are first separated by their motions and, with the evolution of a memory for shapes, form vision progressively evolves. The importance of these studies on lower animals becomes very clear when we take into account the view commonly held by leaders in this field, that the principles governing visual motor control are basically the same in lower animals and humans—whereas, of course, we humans and other primates can see without relative motion between ourselves and our surroundings.

In the last decades the part of the brain in primates responsible for visual processing—the visual cortex—has been studied from an anatomical, physiological, and also behavioral viewpoint. The different parts of the visual cortex have been identified and most of their connections established.

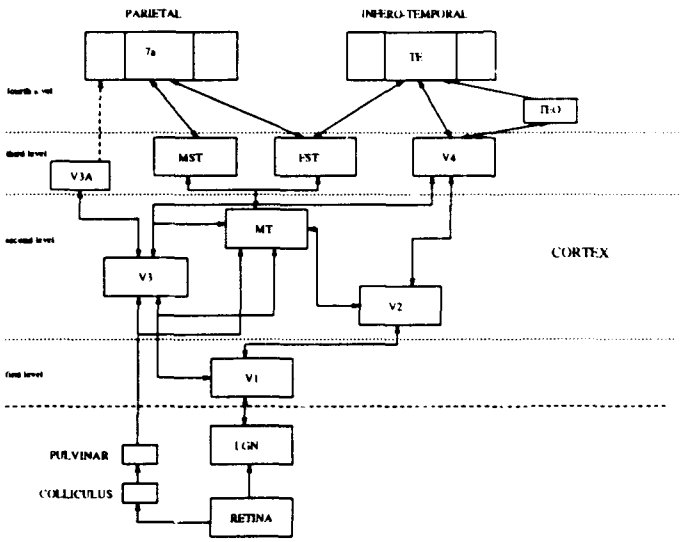

FIGURE 1.4. Diagram of the primate visual system indicating the subcortical structure as well as the four tentative levels of cortical visual processing (from (Orban, 1992)).

Most scientists subscribe to the theory that the different parts perform functionally specialized operations. What exactly these functions are, has not yet been clarified. In particular, opinions diverge about the specialization and the interconnections involved in later stages of processing of the visual data. Much more is known about the earlier processes. The visual signal reaches the cortex at the primary visual cortex—also called V1, or striate cortex, via the retina and the lateral geniculate body. From the primary visual cortex the visual signals are sent to about 30 extrastriate or higher-order visual cortical areas, among which about 300 connections have been reported. Fig. 1.4, taken from (Orban, 1992) shows the major areas involved in visual processing. According to Orban the modules in the primate visual cortex can be divided into four hierarchical levels of processing. It seems to be pretty well accepted that there exist lower areas that are specialized for the processing of either static or dynamic imagery. MT (also called V5), MST, and FST seem to be involved in motion processing, and V4 in color processing. Form vision seems to be accomplished by different lower modules, which use both static and dynamic information. Zeki (1993), for example, suggests that V3 is responsible for the understanding of form from motion information, and V4 derives form and color information. At later stages the modules process both kinds of information in a combined way.

On the basis of anatomical evidence and behavioral studies (studies on patients with lesions of specific cortical areas) the hypothesis has been

brought forward (Ungerleider & Mishkin, 1982) that there exist two visual pathways originating from V1; a dorsal one leading to the parietal cortex and a ventral one leading to the infero-temporal cortex. The dorsal pathway is concerned with either the computations concerned with "where" (object localization) or "how" (the visual guidance of movements (Goodale, Milner, Jacobson & Carey, 1991)), and the ventral pathway with the computations concerned with "what" (object identification). It would be an oversimplification to conceive of these two pathways as being mutually exclusive and hierarchically organized (Zeki, 1993); one of the reasons is that this theory needs to provide an answer to where and how the knowledge of "what" an object is might be integrated with the knowledge of "where" it is. Also, recently the existence of a third pathway leading to the identification of actions has been suggested (Boussaud, Ungerleider & DeSimone, 1990).

Results from the brain sciences show us that there doesn't exist just one hierarchy of visual processes, but various different computations are performed in parallel. Also, it isn't our intention to propose one strict hierarchy for developing visual competences. We merely suggest studying competences by investigating more and more complex models, and base more complicated competences on simpler ones. Naturally, it follows that computations concerned with different cues and representations can and should be studied in parallel.

Motion, shape and space competences

If we follow the results from the natural sciences it becomes clear that the most fundamental competences are the ones that involve visual motion. This leads us to the problems of navigation. The competences we encounter in visual navigation encompass representations of different forms. To elucidate the synthetic approach, in the next section we will discuss a series of competences of increasing complexity employing representations of motion, shape, and space. In the following section we will then outline our realizations of the most basic competences in visual navigation, which only require motion information.

Next in the hierarchy follow capabilities related to the understanding of form and shape and the learning of space. Concerning form and shape, our view is that we should not try to adopt the classical idea of computing representations that capture the 3-D world metrically. Psychological studies on the role of the eye movements suggest that fixations play an important role in our understanding of space. It seems that the level on which information from successive fixations is integrated is relatively abstract and that the representations from which organisms operate on the world is 3-D only locally. Therefore, it will be necessary to study new forms of shape representations. In nature too, there doesn't exist just one method of shape representation. As results from Neurobiology show, form perception in human brains takes place in more than just one part of the cortex and is

realized with different kinds of hardware.

Space is also understood from the processing of various cues in a variety of ways. Furthermore, different tasks will require representations of space with regard to different reference systems—not just one, as often has been debated in the past. Representations might be object-centered, ego-centered, or action-driven.

Actions can be very typical for objects. Early perceptual studies have shown that humans are able to interpret moving scenes correctly, even when the static view does not contain any information about the structure. In the experiments of Johansson (1973) subjects were able to recognize animals, as well as specific human beings, given only the motions of light bulbs mounted on the object's joints. Since our viewpoint is that we should formulate competences as recognition procedures, the study of navigation also leads us to the study of action-driven visual processing.

1.4.3 A Hierarchy of Models for Navigational Competences

Navigation, in general, refers to the performance of sensory mediated movement, and visual navigation is defined as the process of motion control based on an analysis of images. A system with navigational capabilities interacts adaptively with its environment. The movement of the system is governed by sensory feedback which allows it to adapt to variations in the environment. By this definition visual navigation comprises the problem of navigation where a system controls its single components relative to the environment and relative to each other.

Visual navigation encompasses a wide range of perceptual competences, including tasks that every biological species possesses such as motion segmentation and kinetic stabilization (the ability of a single compact sensor to understand and control its own motion), as well as advanced specific hand-eye coordination and servoing tasks.

To explain the principles of the synthetic approach, we describe six such competences, all of which are concerned only with the movement of a single compact sensor. These are: egomotion estimation, partial object-motion estimation, independent motion detection, obstacle avoidance, target pursuit, and homing. These particular competences allow us to demonstrate a hierarchy of models concerned with the representation of motion, form and shape. Table 1.1 describes these competences and formulates them as recognition procedures that rely on increasingly more complex models.

In the past, navigational tasks, since they inherently involve metric relationships between the observer and the environment, have been considered as subproblems of the general "structure-from-motion" problem (Ullman, 1979). The idea was to recover the relative 3-D motion and the structure of the scene in view from a given sequence of images taken by an observer

in motion relative to its environment. Indeed, if structure and motion can be computed, then various subsets of the computed parameters provide sufficient information to solve many practical navigational tasks. However, although a great deal of effort has been spent on the subject, the problem of structure from motion still remains unsolved for all practical purposes. The main reason for this is that the problem is ill-posed, in the sense that its solution does not continuously depend on the input.

TABLE 1.1.

egomotion estimation	Recognizing locations of intersection of axis of rotation and axis of translation with image plane by locating patterns on the flow field. Rigid motion model applied globally.
object-motion estimation	Recognition of tracking acceleration. Rigid motion model applied locally.
independent motion detection	Recognition of locations whose flow vectors do not originate from rigid motion. Various motion models responding to nonrigidity.
obstacle avoidance	Recognition of locations that represent parts of the 3-D world on a collision course with observer. Models for time-to-contact.
target pursuit	Recognizing target's location along with label sufficent to plan pursuing. Models of operational space and the motion of the target.
homing	Recognition of routes connecting different locations. Models of shape, form and space.

The most simple navigational competence, according to our definition, is the estimation of egomotion. The observer's sensory apparatus (eye/camera), independent of the observer's body motion, is compact and rigid and thus moves rigidly with respect to a static environment. As we will demonstrate, the estimation of an observer's motion can indeed be based on only the rigid motion model. A geometric analysis of motion fields reveals that the rigid motion parameters manifest themselves in the form of patterns defined on partial components of the motion fields (Fermüller, 1993). Algorithmically speaking, the estimation of motion thus can be performed through pattern recognition techniques.

Another competence, the estimation of partial information about an object's motion (its direction of translation), can be based on the same model; but, whereas for the estimation of egomotion the rigid motion model could

be employed globally, for this competence only local measurements can legitimately be employed. Following our philosophy about the study of perception, it makes perfect sense to define such a competence which appears very restricted. Since our goal is to study visual problems in the form of modules directly related to the visual task in which the observer is engaged, we argue that in many cases when an object is moving in an unrestricted manner (translation and rotation) in the 3-D world we are only interested in the object's translational component, which can be extracted using dynamic fixation (Fermüller & Aloimonos, 1992).

Next in the hierarchy follow the capabilities of independent motion detection and obstacle avoidance. Although the detection of independent motion seems to be a very primitive task, it can easily be shown by a counterexample that in the general case it cannot be solved without any knowledge of the system's own motion. Imagine a moving system that takes an image showing two areas of different rigid motion. From this image alone, it is not decidable which area corresponds to the static environment and which to an independently moving object.

However, such an example shouldn't discourage us and drive us to the conclusion that egomotion estimation and independent-motion detection are "chicken and egg" problems, that unless one of them has been solved, the other can't be addressed either. Have you ever experienced the illusion that you are sitting in front of a wall which covers most of your visual field, and suddenly this wall (which actually isn't one) starts to move? You seem to experience yourself moving. It seems that vision alone does not provide us (humans) with an infallible capability of estimating motion. In nature the capability of independent motion detection appears at various levels of complexity. We argue that in order to achieve a very sophisticated mechanism for independent motion detection, various different processes have to be employed. Another glimpse at nature should give us some inspiration: We humans also do not perceive everything moving independently in our visual field. We usually concentrate our attention on the moving objects in the center of the visual field (where the image is sensed with high resolution) and pay attention only if something is moving fast in the periphery. It thus seems to make sense to develop processes that detect anything moving very fast (Nelson, 1991). If some upper bound on the observer's motion is known (maximal speed), it is possible to detect even for small areas where motions above the speed threshold appear. Similarly, for specific systems, processes that recognize specific types of motion may be devised by employing filters that respond to these motions (of use, for example, when the enemy moves in a particular way). To cope with the "chicken and egg" problem in the detection of larger independently moving objects, we develop a process based on the same principle as the estimation of egomotion, which for an image patch recognizes whether the motion field within the patch originates from only rigid motion, or whether the constraint of rigidity does not hold. Having some idea about the egomotion or the scene (for

example, in the form of bounds on the motion, or knowing that the larger part of the scene is static) we can also decide where the independently moving objects are.

In order to perform obstacle avoidance it is necessary to have some representation of space. This representation must capture in some form the change of distance between the observer and the scene points which have the potential of lying in the observer's path. An observer that wants to avoid obstacles must be able to change its motion in a controlled way and must therefore be able to determine its own motion and set it to known values. As can be seen, the capability of egomotion estimation is a prerequisite for obstacle avoidance mechanisms, and general independent motion detection will require a model which is as complex as that used in egomotion estimation in addition to other simple motion models.

Even higher in the hierarchy are the capabilities of target pursuit and homing (the ability of a system to find a particular location in its environment). Obviously, a system that possesses these capabilities must be able to compute its egomotion, and avoid obstacles while detecting independent motion. Furthermore, homing requires knowledge of the space and models of the environment (for example, shape models), whereas target pursuit relies on models for representing the operational space and the motion of the target. These examples should demonstrate the principles of the synthetic approach, which argues for studying increasingly complex visual capabilities and developing robust (qualitative) modules in such a way that more complex capabilities require the existence of simpler ones.

1.4.4 Motion-based Competences

In this section we describe the ideas behind some of the modules we have developed to realize the most basic competences for visual navigation: the competence of egomotion estimation, a process for partial object motion estimation and a process for independent motion detection. This description should merely serve to demonstrate our viewpoint concerning the implementation of qualitative algorithms; more detailed outlines and analyses are found elsewhere.

First, let us state some of the features that characterize our approach to solving the above mentioned competences, and distinguishes it from most existing work.

In the past, the problems of egomotion recovery for an observer moving in a static scene and the recovery of an object's 3-D motion relative to the observer, since they both were considered as reconstruction problems, have been treated in the same way. The rigid motion model is appropriate if only the observer is moving, but it holds only for a restricted subset of moving objects—mainly man-made ones. Indeed, all objects in the natural world move non-rigidly. However, considering only a small patch in the image of a moving object, a rigid motion approximation is legitimate. For the case

of egomotion, data from all parts of the image plane can be used, whereas for object motion only local information can be employed.

Most current motion understanding techniques require the computation of exact image motion (optical flow in the differential case or correspondence of features in the discrete case); however, this amounts to an ill-posed problem and additional assumptions about the scene have to be employed. As a result, in the general case, the computed image displacements are imperfect. In turn, the recovery of 3-D motion from noisy flow fields has turned out to be a problem of extreme sensitivity with small perturbations in the input causing large amounts of error in the motion parameter estimation. To overcome this problem, in our approach to the development of motion related competences, we skip the first computational step. All the techniques developed are based on the use of only the sign of the projection of the motion vector along some directions. That is, we assume that the system has the capability to estimate the direction (positive or negative) of the projection of the motion vector along a set of directions. The minimum a system can accomplish is to estimate the direction of retinal motion in at least one direction, namely the one perpendicular to the local edge, or as is known, the direction of the normal flow. It should be mentioned that a few techniques using normal flow have appeared in the literature; however, they deal with restricted cases (only translation or only rotation (Aloimonos & Brown, 1984; Horn & Weldon, 1987)).

Another characteristic is that the constraints developed for the motion modules, for which the rigid motion module is the correct one globally, are such that the input also is utilized globally. The basis of these computations forms global constraints which relate the spatiotemporal derivatives of the image intensity function to the 3-D motion parameters.

If we consider a spherical retina translating with velocity $\vec{t}$, then the motion field is along the great circles connecting the two antidiametric points, the focus of expansion (FOE) and the focus of contraction (FOC), where the vector $\vec{t}$ intersects the sphere. In the case where the eye rotates with velocity $\vec{w}$, the motion field is along the circles where planes perpendicular to $\vec{w}$ cut the sphere. The points where $\vec{\omega}$ cuts the sphere are denoted as AOR and $-$AOR. In the case of rigid motion (the retinas of all moving organisms undergo rigid motion, even if the organisms themselves move nonrigidly) the motion field is the addition of a translational field and a rotational field. In this case it is not easy to recognize the FOE and the AOR; however, if we examine the projection of the motion field along a set of directions we discover a rich global structure. These directions are defined below as the longitudinal and latitudinal vector fields (Figs. 1.5(A) and (B)).

Consider an axis $\vec{s}$ passing from the center of the sphere and cutting the sphere at points N and S. The unit vectors tangential to the great circles containing $\vec{s}$ define a direction for every point on the retina (Fig. 1.5(A)). We call these directions $\vec{s}$-longitudinal, as they depend on the axis $\vec{s}$. Sim-

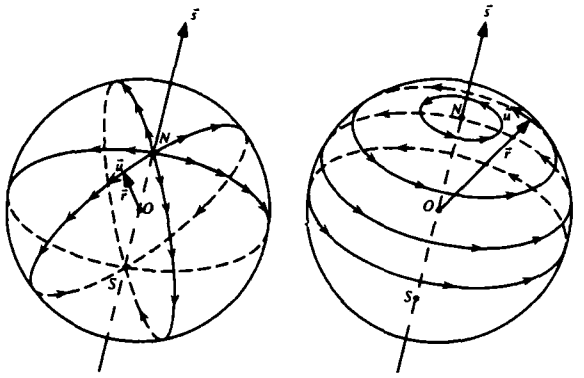

FIGURE 1.5. (A) An axis $\vec{s}$ passing from the center of the sphere and cutting the sphere at points S and N defines a longitudinal vector field. At each point we consider the unit vector tangent to the geodesic connecting S and N. The value of the vector $\vec{u}$ at point $\vec{r}$ is: $u = \frac{\vec{s}-(\vec{s}\cdot\vec{r})\vec{r}}{\|\vec{s}-(\vec{s}\cdot\vec{r})\vec{r}\|}$. (B) An axis $\vec{s}$ passing from the center of the sphere and cutting the sphere at points S and N defines a latitudinal vector field. At each point we consider the unit vector tangent to the circle which is the intersection of the sphere with the plane perpendicular to $\vec{s}$. The value of vector $\vec{u}$ at point $\vec{r}$ is $\vec{u} = \frac{(\vec{s}\times\vec{r})}{\|\vec{s}\times\vec{r}\|}$.

ilarly, we define the $\vec{s}$-latitudinal directions as the unit vectors tangential to the circles resulting from the intersection of the sphere with planes perpendicular to $\vec{s}$ (Fig. 1.5(B)). In the case of a planar retina the longitudinal and latitudinal vector fields become as in Figs. 1.6(A) and (B).

We introduce here a property of these directions that will be of use later. Consider two axes $\vec{s}_1$ and $\vec{s}_2$ cutting the sphere at N_1, S_1 and N_2,S_2 respectively. Each axis defines on every point a longitudinal and a latitudinal direction. We ask the question: where on the sphere are the $\vec{s}_1$ longitudinal (or latitudinal) directions perpendicular to the $\vec{s}_2$ longitudinal (or latitudinal) directions? Considering the $\vec{s}_1$ and $\vec{s}_2$ longitudinal (or latitudinal) directions this question translates to: where on the sphere a great circle containing $\vec{s}_1$ will be perpendicular to a great circle containing $\vec{s}_2$? The answer is in general two closed curves on the sphere defined by the equation $(\vec{r}.\vec{s}_1)(\vec{r}.\vec{s}_2) = \vec{s}_1.\vec{s}_2$, where $\vec{r}$ denotes position on the sphere. The geometry of these curves is described in Fig. 1.7. Considering now the longitudinal directions of one axis and the latitudinal directions of the other axis, they are perpendicular to each other along the great circle defined by the axes $\vec{s}_1$ and $\vec{s}_2$. (Fig. 1.8).

We now would like to examine the structure of the projection of a rigid motion field on an $\vec{s}\,(NS)$ longitudinal set of directions. Since a rigid motion

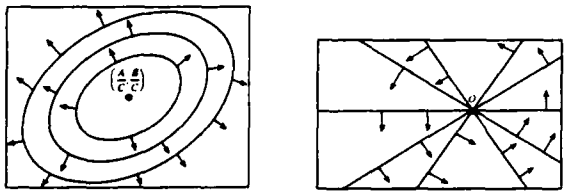

FIGURE 1.6. (A) For the case of a planar retina, the longitudinal field becomes as in the figure, with the vectors perpendicular to a set of conic sections, defined below. An axis $\vec{s} = (A, B, C)$ passing from the nodal point of the eye cuts the image plane at the point $(\frac{A}{C}, \frac{B}{C})$. The family of cones with $\vec{s}$ as their axis intersects the image plane at the set of conics. We have called this field the "co-axis" field (as it is defined by an axis). (B) For the case of a planar retina the latitudinal field becomes as in the figure, with the vectors perpendicular to lines passing from a single point O, which defines the field. We have called this field the "co-point" field.

field is the addition of a translational and a rotational field, we first study the cases of pure translation and pure rotation.

If we project a translational motion field on the $\vec{s}$ longitudinal vectors, the resulting vectors will either be zero, positive (pointing towards S) or negative (pointing towards N). The vectors will be zero on two curves (symmetric around the center of the sphere) whose shape depends on the angle between the vectors $\vec{t}$ and $\vec{s}$ as in Fig. 1.7. Inside the curves the vectors will be negative and outside the curves positive (Fig. 1.9).

If we project a rotational motion field on the $\vec{s}$ (NS) longitudinal vectors, the projections will be either zero (on the great circle defined by $\vec{\omega}$ and $\vec{s}$), positive (in the one hemisphere) or negative (in the other hemisphere) (Fig. 1.10).

If the observer now translates and rotates with velocities $\vec{t}$ and $\vec{\omega}$ it is possible to classify some parts of the projection of the general motion field on any set of $\vec{s}$ longitudinal vectors by intersecting the patterns of Figs. 1.9 and 1.10. If at a longitudinal vector the projection of both the translational and rotational vectors is positive, then the projection of the image motion vector (the sum of the translational and rotational vectors) will also be positive. Similarly, if the projections of both the translational and rotational vectors on a longitudinal vector at a point are negative, so also will the projection of the motion vector at this point. In other words, if we intersect the patterns of Figs. 1.9 and 1.10, whenever positive and positive come together the result will be positive and whenever negative and negative come together the result will be negative. However, whenever positive and negative come together, the result cannot be determined without knowledge of the environment.

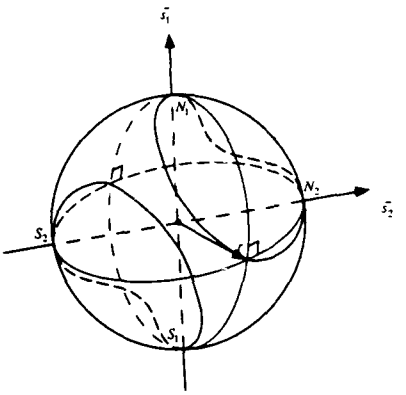

FIGURE 1.7. The great circles containing $\vec{s}_1$ and $\vec{s}_2$ are perpendicular at points of the sphere lying on two closed curves. If $\vec{r}$ denotes a point on the curves, then $(\vec{s}_1 \cdot \vec{r})(\vec{s}_2 \cdot \vec{r}) = \vec{s}_1 \cdot \vec{s}_2$. The shape of the curves depends on the angle between $\vec{s}_1$ and $\vec{s}_2$.

Thus, if we project a rigid motion field on an $\vec{s}$ longitudinal vector field, then the projections will be strictly negative or strictly positive in the areas identified in Fig. 1.11. In the rest of the sphere the projections can be negative, positive or zero. The pattern of Fig. 1.11 is defined by one great circle containing $\vec{\omega}$ and $\vec{s}$ and by two curves containing the points FOE, FOC, N and S.

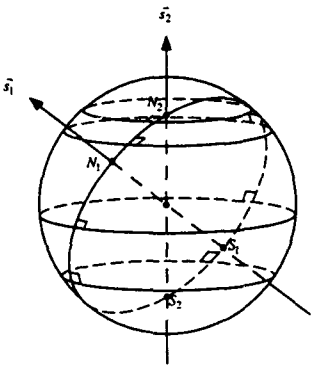

FIGURE 1.8. The $\vec{s}_1$-longitudinal vectors are perpendicular to the $\vec{s}_2$-latitudinal vectors along the great circle defined by $\vec{s}_1$ and $\vec{s}_2$.

It is worth pointing out that the pattern of Fig. 1.11 depends only on the directions of vectors $\vec{s}$ (that defines the longitudinal vectors), $\vec{t}$ and $\vec{\omega}$; and is independent of the scene in view. Also, the pattern is different for a different choice of the vector $\vec{s}$.

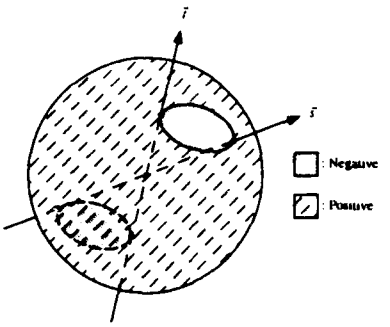

FIGURE 1.9. Projection of a translational motion field on an $\vec{s}$ longitudinal pattern. It is zero on two curves on the sphere (symmetric with regard to the center of the sphere). The points where $\vec{t}$, $\vec{s}$, $-\vec{t}$ and $-\vec{s}$ intersect the sphere lie on the curves. The values are negative inside the curves and positive outside them.

If we consider the projection of a rigid motion field on the $\vec{s}$ latitudinal directions (defined by the vector $\vec{s}\,(NS)$), we obtain a pattern which is dual to the one of Fig. 1.11. This time, the translational flow is separated into positive and negative by a great circle and the rotational flow by two closed curves passing from the points AOR, $-$AOR, N and S, as in Fig. 1.7.

The geometric analysis described above allows us to formulate the problem of egomotion estimation as a pattern recognition problem. If the system has the capability of estimating the sign of the retinal motion along a set of directions at each point, then this means that the system can find the sign of the longitudinal and latitudinal vectors for a set of axes $\vec{s}_i, i = 1, \ldots, n$. If the system can now locate the patterns in each longitudinal and latitudinal vector field, then it has effectively recognized the directions $\vec{t}$ and $\vec{\omega}$. If, however, the system has less power and can only compute the motion in at most one direction in every point, namely the one perpendicular to the local edge, then the solution proceeds exactly as before. The difference is that for each longitudinal or latitudinal set of directions we do not have information (positive, negative or zero) at every point of the sphere.

Considering a planar retina instead of a spherical retina, we have the co-point vectors instead of the latitudinal vectors and the co-axis vectors instead of the longitudinal vectors (Fermüller, 1993). The curves separating the positive from the negative values become a second order curve and a line in the plane. Fig. 1.12 shows for the case of a planar retina the pattern for the co-point vectors (latitudinal).

Thus we see that utilizing the geometry of the motion field globally, we can get a lot of information from only a part of the image: the part where we know that the vectors are only negative or only positive. Recall that in order to find the pattern of Fig. 1.11, we had to intersect the patterns of Figs. 1.9 and 1.10. At the intersection of positive and negative parts, the

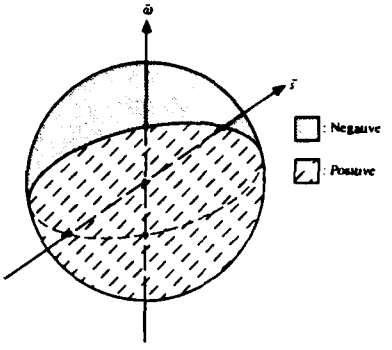

FIGURE 1.10. Projection of a rotational motion field on an $\vec{s}$ longitudinal pattern. The values are zero on the great circle defined by the plane of $\vec{\omega}$ and $\vec{s}$. In the one hemisphere the values are positive and in the other they are negative.

sign depends on the depth. It is only in these areas that the value along the longitudinal or latitudinal vectors can become zero. The distribution of the image points where the normal flow in some direction becomes zero has again a rich geometric structure containing egomotion information. The interested reader is referred to (Fermüller & Aloimonos, 1994).

Finally, based on the same basic constraints, a process for the detection of independent motion has been designed. Since the observer is moving rigidly, an area with a motion field not due to only one rigid motion, must contain an independently moving object. The constraints are defined for the whole visual field, but also the motion vectors in every part of the image plane must obey a certain structure. Our approach consists of comparing the motion field within image patches with pre-stored patterns (which represent all possible rigid motions).

By considering patches of different sizes and using various resolutions, the patterns may also be of use in estimating the motion of objects. Differently sized filters can first be employed to localize the object and then an appropriately sized filter can be used to estimate the motion; however, objects do not always move rigidly. Furthermore, in many cases the area covered by the object will not be large enough to provide satisfyingly accurate information. In the general case, when estimating an object's motion, only local information can be employed. In such a case, we utilize the observer's capability to move in a controlled way. We describe the object's motion with regard to an object centered coordinate system. From fixation on a small area on the object the observer can derive information about the direction of the object's translation parallel to its image plane. By tracking the object over a small amount of time, the observer derives additional information about the translation perpendicular to the image plane. Combining the computed values allows the observer to derive the direction of

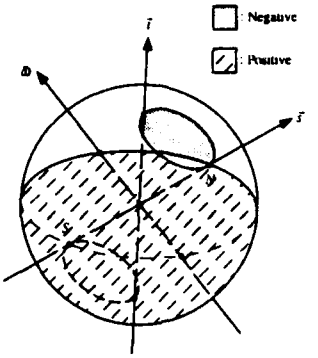

FIGURE 1.11. The projection of a rigid motion field on an $\vec{s}$ longitudinal pattern. The sphere is divided in two halves with the great circle of the plane defined by $\vec{\omega}$ and $\vec{s}$. There are also two curves (the ones of Fig. 1.9) passing from the points where $\vec{t}$, $\vec{s}$, $-\vec{t}$ and $-\vec{s}$ intersect the sphere. Whatever the motion $\vec{t}$ and $\vec{\omega}$ is, there exists a pattern of positive and negative longitudinal vectors in a part of the sphere. (The intersection of the negative parts of Figs. 1.9 and 1.10 provides the negative part and the intersection of the positive parts provides the positive.)

an object's translation (Fermüller & Aloimonos, 1993).

1.4.5 A LOOK AT THE MOTION PATHWAY

There is a very large amount of literature (Duffy & Wurtz, 1991; Maunsell & Essen, 1983; Tanaka & Saito, 1989; Ungerleider & DeSimone, 1986) on the properties of neurons involved in motion analysis. The modules which have been found to be involved in the early stages of motion analysis are the retinal parvocellular neurons, the magnocellular neurons in the LGN, layer 4Cβ of V1, layer 4B of V1, the thick bands of V2 and MT. These elements together are referred to as the early motion pathway. Among others they feed further motion processing modules, namely MST and FST, which in turn have connections to the parietal lobe. Here we present a hypothesis, based on the computational model described earlier, about how motion is handled in the cortex.

Fig. 1.13 (from (Movshon, 1990)) shows an outline of the process to be explained which involves four kinds of cells with different properties. In the early stages, from the retinal Pa ganglion cells through the magnocellular LGN cells to layer 4Ca of V1 the cells appear functionally homogeneous and respond almost equally well to the movement of a bar (moving perpendicularly to its direction) in any direction (Fig. 1.13(A)). Within layer 4C of V1 we observe an onset of directional selectivity. The receptive fields of the neurons here are divided into separate excitatory and inhibitory regions. The regions are arranged in parallel stripes and this arrangement

provides the neurons with a preference for a particular orientation of a bar target (which is displayed in the polar diagram) (Fig. 1.13(B)). In layer 4B of V1 another major transformation takes place with the appearance of directional selectivity. The receptive fields here are relatively large and they seem to be excited everywhere by light or dark targets. In addition, these neurons respond better or solely to one direction of motion of an optimally oriented bar target, and less or not at all to the other (Fig. 1.13(C)). Finally, in MT neurons have considerably large receptive fields and in general the precision of the selectivity for direction of motion that the neurons exhibit is typically less than in V1 (Fig. 1.13(D)).

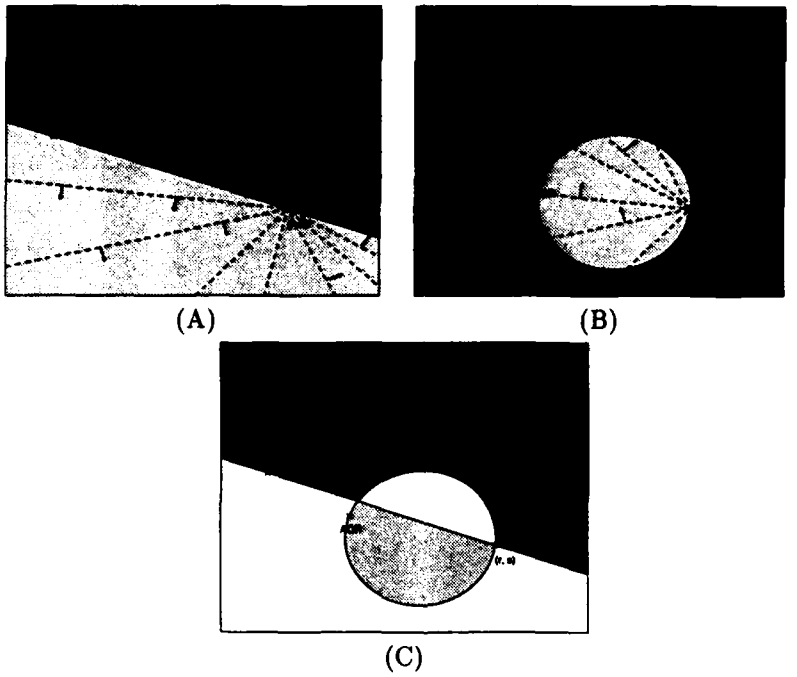

FIGURE 1.12. (A) The translational (r, s) co-point vectors are separated by a line that passes through the FOE (the point which denotes the direction of translation); in one half-plane all vectors have positive values (light grey), in the other half-plane negative values (dark grey). (B) The rotational (r, s) co-point vectors are separated by a second order curve that passes through the AOR (the point where the rotation axis pierces the image plane). (C) A general rigid motion separates the (r, s) co-point vectors into an area of negative vectors, an area of positive vectors, and an area that may contain vectors of any value (white).

One can easily envision an architecture that, using neurons with the properties listed above implements a global decomposition of the normal

motion field: Neurons of the first kind could be involved in the estimation of the local retinal motion perpendicular to the local edge (normal flow). Neurons at this stage could be thought of as computing whether the projection of retinal motion along some direction is positive or negative. Neurons of the second kind could be involved in the selection of local vectors in particular directions as parts of the various different patterns discussed in the previous section, while neurons of the third kind could be involved in computing the sign (positive or negative) of pattern vectors for areas in the image; i.e., they might compute patches of different sizes whether the normal flow in certain directions is positive or negative. Finally, neurons of the last kind could be the ones that piece together the parts of the patterns developed already into global patterns that are matched with prestored global patterns. Matches provide information about egomotion and mismatches provideinformation about independent motion.

In this architecture we are not concerned with neurons that possibly estimate the motion field (optic flow). This is not to say that optic flow is not estimated in the cortex; several neurons could be involved in approximating the motion field. However, if the cortex is capable of solving some motion problems without the use of optic flow, whose estimation amounts to the solution of an optimization problem, it is quite plausible to expect that it would prefer such a solution. After all, it is important to realize that at the low levels of processing the system must utilize very reliable data, such as the sign of the motion field along some direction. It is worth noting that after deriving egomotion from normal flow, information about 3-D motion is available, and the cortex could involve itself with approximating optic flow, because in this way the problem is not ill-posed any more (at least for background scene points).

1.4.6 Form-based Competences

Since Computer Vision was considered to be approached through the construction of 3-D descriptions of the world, a lot of effort was spent on developing techniques for computing metric shape and depth descriptions from 2-D imagery. Studies concerned with this kind of work are collectively referred to as "shape from X" computations, where X refers to cues such as shading, texture, pattern, motion, or stereo. Exact, quantitative 3-D structure is hard to compute though, and explicit assumptions about the scene (smoothness, planarity, etc.) usually have to be made in the models employed.

Considering all the work that has been spent on the computation of metric shape and that has yet not given rise to any system working in a real environment, a glimpse at nature might give us some inspiration. Maybe it is a hopeless task to aim at deriving metric shape or depth information. Psychophysical experiments indicate that binocular stereopsis in the human visual system does not produce an explicit representation of

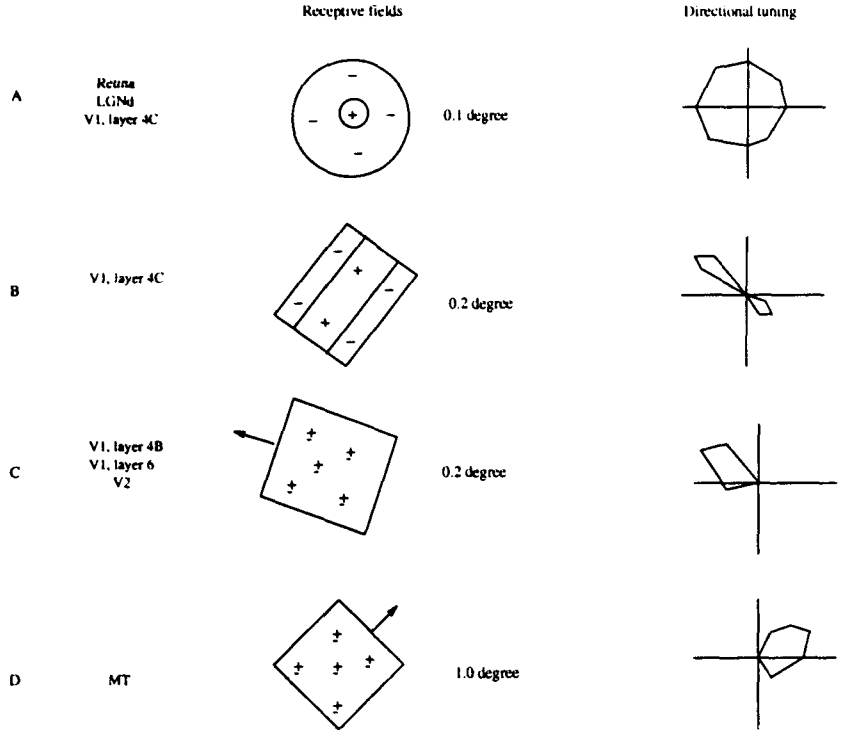

FIGURE 1.13. The spatial structure of visual receptive fields and their directional selectivity at different levels of the motion pathway (from (Movshon, 1990)). The spatial scales of the receptive fields (0.1 degree, etc.) listed here are for neurons at the center of gaze; in the periphery these dimensions would be larger. The polar diagrams illustrate responses to variation in the direction of a bar target oriented at right angles to its direction of motion. The angular coordinate in the polar diagram indicates the direction of motion and the radial coordinate the magnitude of the response.

the metric depth structure of the scene. Psychophysical evidence (Collett, Schwartz & Sobel, 1991; Johnston, 1991) suggests that human performance in tasks involving metric structure from binocular disparities is very poor. Also, other cues don't seem to allow humans to extract the kind of depth information that has usually been considered. In their experiments, Todd and Reichel (1989) had subjects estimate the depths of points on a drape-like surface shown on video images. Subjects could accurately report the relative depth of two points if they were on the same surface on the same side of the "fold," but were quite poor at determining the relative depth if the points were on different "folds." This experiment leads to the conclusion that humans possess relative depth judgment for points within a local area lying on a surface; however, they cannot estimate even relative

depth correctly for large distances in the visual field, when depth extrema are passed.

We also know that in humans the area of the eye in which detailed (high resolution) information can be extracted covers only a small region around the fovea (about five degrees of visual angle at normal viewing distance). The low resolution at the periphery does not allow to derive accurate depth information. Human eyes, however, are seldom not in motion. The eyes are engaged in performing fixations, each lasting about 1/4 of a second. Between the fixations, saccadic movements are carried out, during which no useful information is extracted.

The biological evidence gives us good reason to argue for alternative shape models. The experiments mentioned above give rise to the following conclusions:

(a) Shape or depth should not be computed in metric form, but only relative depth measurements (ordered depth) should be computed.

(b) Shape/depth information should be computed only locally. Then the information derived for different patches has to be integrated. This integration, however, should not take place in the usual form, leading to complete, coherent spatial descriptions. The result should not be a complete reconstructed 3-D shape model, obtained by exactly putting ("glueing") together the local shape representations to a global one. Instead, we have to look for alternative representations that suffice for accessing the shape information one needs to solve particular tasks.

These or similar arguments also find support from computational considerations. Concerning argument (b), one might ask why one should compute only local information, if from a technical standpoint there is no difference whether the devised sensors have different or the same resolution everywhere. If stereo systems are used—the most obvious for deriving shape information—and the two cameras fixate at a point, the disparity measurements are small only near the fixation point, and thus can also only be computed exactly there. In particular, if continuous techniques are employed to estimate the displacement (due to stereo or also due to motion), the assumption of continuity of the spatio-temporal imagery does not have to be greatly violated. The measurements which are due to rotation increase with the distance from the image center and the translational measurements are proportional to the distance from the epipole or the point denoting the direction of translation. Another argument is that computing shape only locally gives legitimacy to the the orthographic projection model for approximating the image formation. The exact perspective projection model makes the computation of distance and shape very hard, since the depth component appears inversely in the image coordinates, which in turn leads to equations that are non-linear in the unknown parameters.

However, concerning argument (a), we don't want to prescribe to the computation of ordered as opposed to metric shape information. Why

should we limit ourselves to ordered depth and not be even less restrictive? Throughout this chapter, we have argued for task-dependent descriptions. This also applies to the shape descriptions; a variety of shape descriptions subserving different tasks can be accepted. To derive metric depth or shape means to compute exact values of the distance between the camera and the scene. In order to solve, for example, the general structure from motion problem, theoretically we require at least three views of the scene, or two views and some additional information, such as the length of the baseline for a stereo setting. From two perspective views, only scaled distance, or distance up to the so-called relief transformation, can be derived. To compute only ordered depth measurements would mean that in addition, scaled depth is derived only up to a positive term (i.e., it would result in deriving monotonic functions of the depth measurement Z, for example functions of the form $f(Z) = \frac{1}{Z}a + b$, $f(Z) = aZ + b$, etc., where a and b are constants). We then argue that one could try to compute even less informative depth or shape information by aiming at deriving more involved depth functions.

Under the influence of the reconstructionists' ideas, all effort in the past has been devoted to deriving metric measurements. A new look at the old research with a different goal in mind might give us new insights. From different cues, depth and shape information of different forms might be computed and then appropriately fused. A representation less than an ordered one by itself does not seem to be sufficient for 3-D scene understanding. However, by combining two or more such representations, additional information can be obtained. It seems that the study of fusion of information for the purpose of deriving form and shape description will definitely be of importance.

It should be noted that whereas shape and depth measurements are equivalent for a metric 3-D representation, they are not for ordered representations. Dealing with metric measurements, if absolute depth is given, shape (defined as the first order derivatives of depth) can be directly computed, and vice versa. The same, however, does not hold for ordered, or even less informative representations.

Our goal is to derive qualitative, as opposed to quantitative, representations, because the computations to be performed should be robust. This requires that we not make unreasonable assumptions and employ computations that are ill-posed. Qualitativeness, for example, does not mean performing the same computations that have been performed under the reconstruction philosophy, making the same assumptions about the 3-D world, and at the end separating the computed values by a threshold in order to end up with "qualitative" information in the form of "greater or smaller than some value." Our effort should be devoted to deriving qualitative shape descriptions from well-defined input. For example, it wouldn't make sense to assume exact optical flow or stereo disparity measurements—which are impossible to obtain—in order to derive shape descriptions less powerful than the one of scaled depth because, if we had exact 2-D image

measurements, we could compute scaled shape, and there is nothing we would gain computationally from computing less.

By concentrating on simpler shape descriptions, new mathematical models and new constraints might be found. Purely mathematical considerations can reveal what kind of information could possibly be computed from a certain input allowing a defined class of operations. The study of Koenderink and van Doorn (1991) on affine structure from motion might serve as an inspiration; in it they investigated a hierarchy of shape descriptions based on a stratification of geometries.

1.4.7 SPACE UNDERSTANDING

Since in the past the actions of the observer were not considered as an integral part of perceptual investigations, computational modelling, and in particular AI research has dealt with space only at a symbolic level. For example, some early systems (Winston, 1975) dealt with the spatial relationship of objects in a blocks world. Assuming that objects can be recognized and thus can be stored as symbols, the spatial configuration of these objects under changing conditions was studied. Also, in existing studies on spatial planning (e.g., path planning), solutions to the problems of recognizing the objects and the environment are assumed to be available for the phase of coordinating motions.

Within the framework of behavioral vision a new meaning is given to the study of space perception. The understanding of the space surrounding an observer results from the actions and perceptions the observer performs and their relationship. For a static observer that does not act in any way, space does not have much relevance. But in order to interact with its environment it has to have some knowledge about the space in which it lives, which it can acquire through actions and perceptions. Of course, the knowledge of space can be of different forms at various levels of complexity depending on the sophistication of the observer/actor and the tasks it has to perform. At one end of the scale, we find a capability as simple as obstacle avoidance, which in the most parsimonious form has to capture only the distance between the observer and points in the 3-D world, and at the other end of the scale, the competence of homing, which requires the actor to maintain some kind of map of its environment.

To obtain an understanding of space by visual means requires us to identify entities of the environment and also to localize their positions; thus both basic problems, the one of "where" and the one of "what" have to be addressed.

The problem of recognizing three-dimensional objects in space is by itself very difficult, since the object's appearance varies with the pose it has relative to the observer. In the Computer Vision literature two extreme views are taken about how to address the 3-D recognition problem, which differ in the nature of the models to be selected for the descriptions of

objects in the 3-D environment. One view calls for object-centered models and the other for descriptions of the objects by means of viewer-centered views (3-D vs 2-D models). In most of the work on object-centered descriptions the form of objects is described with simple geometric 3-D models, such as polyhedra, quadrics, or superquadrics. Such models are suited to represent a small number of man-made (e.g., industrial) parts. However, to extend 3-D modelling to a larger range of objects will require models of more complex structural description, characterizing objects as systems and parts of relations. Recently a number of studies have been performed on viewer-centered descriptions approaching the problem from various directions. To name a few of them: Based on some results in the literature of structure from motion, that show that under parallel projection any view of an object can be constructed as a linear combination of a small number of views of the same object, a series of studies on recognition using orthographic and paraperspective projections have been conducted (Ullman & Basri, 1991; Jacobs, 1992). The body of projective geometry has been investigated to prove results about the computation of structure and motion from a set of views under perspective projection (Faugeras, 1992). The learning of object recognition capabilities has been studied for neuronal networks using nodes that store viewer-centered projections (Poggio, Edelman & Fahle, 1992), and geometric studies on the so-called aspect graphs have investigated how different kinds of geometric properties change with the views the observer has of the geometric model (Koenderink & van Doorn, 1979).

The problem of solving both localization and recognition is exactly the antagonistic conflict at the heart of pattern recognition. From the point of signal processing, it has been proved (Gabor, 1946) that any single (linear) operator can answer only one of these questions with sufficient accuracy. In theory, thus, a number of processes are required to solve tasks related to space perception.

Results from the brain sciences reveal that the receptive field sizes of cells are much larger in the specialized visual areas involved in later processing than in those of the early stages. Many cells with large receptive field sizes respond equally well to stimuli at different positions. For example, in V5 cells with large receptive fields respond to spots of lights moved in certain directions, no matter where the stimulus in the receptive field occurs; nevertheless, the position of the light in the visual field can be localized accurately. Neurobiologists have suggested several solutions to this problem. The following interesting results deserve special mention: In the visual cortex cells have been found which are "gaze-locked," in the sense that they only respond to a certain stimulus if the subject is gazing in a particular direction. These cells probably respond to absolute positions in the ego-centric space (Zeki, 1993).

It seems that nature has invented a number of ways for perceiving space through recognition and localization of objects in the 3-D world. Also,

neurophysiological studies have been conducted that give good reason to assume that the perception of space in primates is not only grounded on object-centered or ego-centered descriptions, but that some descriptions are with regard to some action. For example, in an area called TEA, cells have been reported which are involved in the coding of hand movements (Perrett, Harries, Mistlin & Chitty, 1990). These cells respond when an action is directed towards a particular goal, but they do not respond to the component actions and motions when there is no causal connection between them. Monkeys were shown on video film arrangements of hand movements and object movements contiguous or separated in space or time, for example, a hand and a cup. The hand was retracted and after a short delay the cup moved (as by itself) along the same trajectory as the hand. As the discrepancy between hand and object movement widened the impression of causality weakened. The mentioned cells tuned to hand actions were found to be less responsive when the movement of the hand and the object were spatially separated and appeared not to be causally related.

Humans possess a remarkable capability in recognizing situations, scenes, and objects in the space surrounding them from actions being performed. In the Computer Vision literature a number of experiments (Johansson, 1973) are often cited in which it has been shown that humans can recognize specific animals and humans that move in the dark and are visible only from a set of flashing light bulbs attached to their joints. These experiments demonstrate very well the power of motion cues. Since actions give rise to recognition, and actions are largely understood from motions, it seems to be worthwhile to investigate further motion models, more complicated than the rigid one, to describe actions. For example, situations occurring in manipulation tasks might be modelled through non-rigid motion fields. The change of the motion field or parts of it may be expressed in form of space-time descriptions that can be related to the tasks to be performed. It should be mentioned that recently some effort along this line has started; a few studies have been conducted exploiting motion cues for recognition tasks. In particular, periodic movements, such as the motion of certain animal species have been characterized in frequency space (Nelson & Polana, 1992; Shavit & Jepson, 1993). Statistical pattern recognition techniques have been applied in the time-domain to model highly structured motions occurring in nature, such as the motions of flowing water or fluttering leaves (Polana & Nelson, 1993). Attempts have been made to model walking or running humans by describing the motion of single limbs rigidly (Qian & Huang, 1992), and also various deformable spatial models like superquadrics and snakes have been utilized to model non-rigid motions of rigid bodies (Pentland, Horowitz & Sclaroff, 1991), for the purpose of face recognition.

Representations used for understanding space should be allowed to be of any of three kinds: with regard to the viewer, with regard to an object, or action-driven. An appropriate representation might allow us to solve tasks

straightforwardly that would require very elaborate computations and descriptions otherwise. Perrett et al. (1988) give a good example underpinning this point of view: A choreographer could, for example, use a set of instructions centered on the different dancers (such as to dancer M. who is currently lying prostrate and oriented toward the front of the stage: "Raise head slowly" and to dancer G., currently at the rear of the stage facing stage left: "Turn head to look over left shoulder"). Alternatively the choreographer could give a single instruction to all members of the dance troupe ("Move the head slowly to face the audience"). To allow for the choice of different systems of representation will be a necessity when studying space descriptions. These descriptions, however, must be related in some form. After all, all measurements are taken in a frame fixed to the observer's eye. Thus a great deal of work in space understanding will amount to combining different representations into an ego-centered one.

The competence of homing is considered to be the apogee of spatial behavior. The so amazing homing capabilities of some animals have attracted the attention of researchers for many years. In particular, effort has been spent on investigating the sensory basis of animals' perception; discoveries were made about the use of sensory guidance by sunlight, light patterns in the sky, and moonlight, such as the use of ultraviolet light by ants (Lubbock, 1889) and polarized light by bees (Frisch, 1949). Recently, research has also started on investigations of how particular species organize the spatial information acquired through their motor sequences and sensors (Sandini, Gandolfo, Grosso & Tistarelli, 1993; Srinivasan, Lehrer, Zhang & Horridge, 1989).

Zoologists differentiate between two mechanisms for acquiring orientation: the use of ego-centered and geo-centered systems of reference. Simple animals, like most arthropods, represent spatial information in form of positional information obtained by some kind of route integration relative to their homes. The route consists of path segments each of which takes the animal for a given distance in a given direction. This form of representation related to one point of reference is referred to as an ego-centered representation.[1] More complicated than relying only on information collected en route is the use of geo-centered reference systems where the animal in addition relies on information collected on site (recognition of landmarks) and where it organizes spatial information in a map-based form.

However, research from studies on arthropods (Wehner, 1992; Collett, Dillmann, Giger & Wehner, 1992; Collett, Fry & Wehner, 1993) shows that already in these simple animals, the competence of homing is realized in seemingly any possible way. A large variety of different ways employing

[1] In the Computer Vision literature the term "ego-centered" reference system is used with a different meaning than in Zoology.

combinations of information from action and perception have been discovered. The way the path is stored, the way landmarks are recognized, etc., is different for every species. Not many general concepts can be derived; it seems that the physical realizations are tightly linked to the animal's physiology and overall performance. This has to apply to artificial systems as well. Computations and implementations cannot be separated. Obviously, the more storage capability a system has, the more complex operations it can perform. The number of classes of landmarks that a system can differentiate and the number of actions it can perform will determine the homing capability of a system. Our suggested strategy is to address competences involving space representations (and in particular the homing competence) by synthesizing systems with increasing action and perception capabilities and study the performance of these systems, considering constraints on their memory.

1.5 Conclusions

The study of vision systems in a behavioral framework requires the modelling of observer and world in a synergistic way and the analysis of the interrelationship of action and perception. The role that vision plays in a system that interacts with its environment can be considered as the extraction of representations of the space-time in which the system exists and the establishing of relations between these representations and the system's actions. We have defined a vision system as consisting of a number of representations and processes, or on a more abstract level, as a set of maps which can be classified into three categories: the visual competences that map different representations of space-time (including the retinotopic ones) to each other, the action routines which map space-time representations to motor commands or representations of various kinds residing in memory, and the learning programs that are responsible for the development of any map. To design or analyze a vision system amounts to understanding the mappings involved. In this paper we have provided a framework for developing vision systems in a synthetic manner, and have discussed a number of problems concerning the development of competences, learning routines and the integration of action and perception. We have also described some of our technical work on the development of specific motion-related competences.

To achieve an understanding of vision will require efforts from various disciplines. We have described in this study work from a number of sciences, computational as well as empirical ones. Besides these, the general area of Information Processing has various fields of study from which the design and analysis of vision systems can benefit. Some studies of possible interest include the realization of specific maps in hardware (VLSI chips or optical computing elements); the study of the complexity of visual tasks under the

new framework; information theoretic studies investigating the relationship between memory and task-specific perceptual information; and the study of control mechanisms for behavioral systems.

Acknowledgments: The support of ONR, NSF and ARPA is gratefully acknowledged.

1.6 References

Aloimonos, J. & Brown, C. (1984). Direct processing of curvilinear sensor motion from a sequence of perspective images. In *Proc. Workshop on Computer Vision: Representation and Control*, (pp. 72–77).

Aloimonos, J. & Shulman, D. (1993). *Integration of Visual Modules: An Extension of the Marr Paradigm.* Boston: Academic Press.

Aloimonos, J., Weiss, I. & Bandopadhay, A. (1988). Active vision. *International Journal of Computer Vision, 2*, 333–356.

Aloimonos, J. Y. (1990). Purposive and qualitative active vision. In *Proc. DARPA Image Understanding Workshop*, (pp. 816–828).

Aloimonos, Y. (Ed.). (1993). *Active Perception.* Advances in Computer Vision. Hillsdale, NJ: Lawrence Erlbaum.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE, 76*, 996–1005.

Ballard, D. (1991). Animate vision. *Artificial Intelligence, 48*, 57–86.

Boussaud, D., Ungerleider, L. & DeSimone, R. (1990). Pathways for motion analysis: cortical connections of the medial superior temporal fundus of the superior temporal visual areas in the macaque monkey. *The Journal of Comparative Neurology, 296*, 462–495.

Brady, M., Hollerbach, J., Johnson, T., Lozano-Perez, T. & Mason, M. (Eds.). (1983). *Robot Motion.* Cambridge, MA: MIT Press.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, 2*, 14–23.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments.* Berkeley: Univ. of California Press.

Collett, T., Dillmann, E., Giger, A. & Wehner, R. (1992). Visual landmarks and route following in desert ants. *Journal of Comparative Physiology A, 170*, 435–442.

Collett, T., Fry, S. & Wehner, R. (1993). Sequence learning by honeybees. *Journal of Comparative Physiology A, 172*, 693–706.

Collett, T., Schwartz, U. & Sobel, E. (1991). The interaction of oculomotor cues and stimulus size in stereoscopic depth constancy. *Perception, 20*, 733–754.

Duffy, C. & Wurtz, R. (1991). Sensitivity of MST neurons to optical flow stimuli I: a continuum of response selectivity to large field stimuli. *Journal of Neurophysiology, 65*, 1329–1345.

Ernst, G. & Newell, A. (1969). *GPS: A Case Study in Generality and Problem Solving.* New York: Academic Press.

Faugeras, O. (1992). *Three Dimensional Computer Vision.* Cambridge, MA: MIT Press.

Fermüller, C. & Aloimonos, Y. (1994). On the geometry of visual correspondence. Technical Report CAR-TR, Center for Automation Research, University of Maryland.

Fermüller, C. (1993). Navigational preliminaries. In Y. Aloimonos (Ed.), *Active Perception*, Advances in Computer Vision. Hillsdale, NJ: Lawrence Erlbaum.

Fermüller, C. & Aloimonos, Y. (1992). Tracking facilitates 3-d motion estimation. *Biological Cybernetics, 67*, 259–268.

Fermüller, C. & Aloimonos, Y. (1993). The role of fixation in visual motion analysis. *International Journal of Computer Vision: Special issue on Active Vision, M. Swain (Ed.), 11(2)*, 165–186.

Frisch, K. (1949). Die Polarisation des Himmelslichts als orientierender Faktor bei den Tänzen der Bienen. *Experientia, 5*, 142–148.

Gabor, D. (1946). Theory of communication. *Journal of IEE 93, part III*, 429–457.

Gelernter, H. (1959). Realization of a geometry theorem-proving machine. In *Information Processing: Proceedings of the International Conference on Information Processing*, UNESCO.

Gibson, J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Goodale, M., Milner, A., Jacobson, L. & Carey, D. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*, 154–156.

Helmholtz, H. v. (1896). *Handbuch der Physiologischen Optik*. Leopold Voss.

Hervé, J. (1993). *Navigational Vision*. PhD thesis, University of Maryland, Computer Vision Laboratory, Center for Automation Research, University of Maryland.

Horn, B. (1986). *Robot Vision*. New York: McGraw Hill.

Horn, B. & Weldon, E. (1987). Computationally efficient methods for recovering translational motion. In *Proc. International Conference on Computer Vision*, (pp. 2–11).

Horridge, G. (1987). The evolution of visual processing and the construction of seeing systems. *Proceedings of the Royal Society, London B, 230*, 279–292.

Horridge, G. (1991). Evolution of visual processing. In J. Cronly-Dillon & R. Gregory (Eds.), *Vision and Visual Dysfunction*. New York: MacMillan.

Hubel, D. & Wiesel, T. (1968). Receptive fields and functional architecture of the monkey striate cortex. *Journal of Physiology, 195*, 215–243.

Jacobs, D. (1992). Space efficient 3d model indexing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 439–444).

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*, 201–211.

Johnston, E. (1991). Systematic distortions of shape from stereopsis. *Vision Research, 31*, 1351–1360.

Kanizsa, G. (1979). *Organization in Vision: Essays on Gestalt Perception*. New York: Praeger.

Kant, I. (1990). *Critique of Pure Reason*. Buffalo NY: Prometheus Books.

Koenderink, J. & van Doorn, A. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics, 32*, 211–216.

Koenderink, J. & van Doorn, A. (1991). Affine structure from motion. *Journal of the Optical Society of America, 8*, 377–385.

Kohler, W. (1947). *Gestalt Psychology*. New York: Liveright.

Lubbock, J. (1889). *On the Senses, Instincts, and Intelligence of Animals with Special Reference to Insects*. London: K. Paul Trench.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Maunsell, J. & Fssen, D. V. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey I. Selectivity for stimulus direction, speed and orientation. *Journal of Neurophysiology*, *49*, 1127–1147.

Movshon, A. (1990). Visual processing of moving images. In H. Barlow, C. Blakemore & M. Weston-Smith (Eds.), *Images and Understanding* (pp. 122–137). Cambridge University Press.

Nalwa, V. (1993). *A Guided Tour of Computer Vision*. Winston.

Nelson, R. (1991). Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, *7*, 33–46.

Nelson, R. & Polana, R. (1992). Qualitative recognition of motion using temporal *CVGIP: Image Understanding*, *1*, 33–46. Special Issue on Purposive, Qualitative, Active Vision, Y. Aloimonos (Ed.).

Nilsson, N. (1980). *Principles of Artificial Intelligence*. Palo Alto, California: Tioga Publishing Co.

Orban, G. (1992). The analysis of motion signals and the nature of processing in the primate visual system. In G. Orban & H.-H. Nagel (Eds.), *Artificial and Biological Vision Systems*, ESPRIT Basic Research Series (pp. 24–57). Springer-Verlag.

Pentland, A. (Ed.). (1986). *From Pixels to Predicates: Recent Advances in Computational and Robot Vision*. Norwood, NJ: Ablex.

Pentland, A., Horowitz, B. & Sclaroff, S. (1991). Non-rigid motion and structure from contour. In *Proc. IEEE Workshop on Visual Motion*, (pp. 288–293).

Perrett, D., Harries, M., Mistlin, A. & Chitty, A. (1990). Three stages in the classification of body movements by visual neurons. In H. Barlow, C. Blakemore & M. Weston-Smith (Eds.), *Images and Understanding* (pp. 94–107). Cambridge University Press.

Perrett, D., Mistlin, A. & Chitty, M. H. A. (1988). Vision and action: The control of grasping. Ablex Pub.

Poggio, T., Edelman, S. & Fahle, M. (1992). Learning of visual modules from examples: A framework for understanding adaptive visual performance. *CVGIP: Image Understanding*, *56*, 22–30. Special Issue on Purposive, Qualitative, Active Vision, Y. Aloimonos (Ed.).

Polana, R. & Nelson, R. (1993). Detecting activities. In *Proc. IEEE Image Understanding Workshop*, (pp. 569–574).

Qian, R. & Huang, T. (1992). Motion analysis of articulated objects. In *Proc. International Conference on Pattern Recognition*, (pp. A220–223).

Sandini, G., Gandolfo, F., Grosso, E. & Tistarelli, M. (1993). Vision during action. In Y. Aloimonos (Ed.), *Active Perception* (pp. 151–190). Hillsdale, NJ: Lawrence Erlbaum.

Shavit, E. & Jepson, A. (1993). Motion using qualitative dynamics. In *Proc. IEEE Workshop on Qualitative Vision.*

Sommer, G. (1994). Architektur und Funktion visueller System . Künstliche Intelligenz, 12. Frühjahrsschule.

Srinivasan, M., Lehrer, M., Zhang, S. & Horridge, G. (1989). How honeybees measure their distance from objects of unknown size. *Journal of Comparative Physiology A, 165*, 605–613.

Tanaka, K. & Saito, H. (1989). Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells illustrated in the dorsal part of the Medial Superior Temporal area of the macaque monkey. *Journal of Neurophysiology, 62*, 626–641.

Thompson, W. & Pong, T.-C. (1990). Detecting moving objects. *International Journal of Computer Vision, 4*, 39–57.

Todd, J. & Reichel (1989). Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychology Review, 96*, 643–657.

Ullman, S. (1979). *The Interpretation of Visual Motion.* MIT Press.

Ullman, S. & Basri, R. (1991). Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 992–1006.

Ungerleider, L. & DeSimone, R. (1986). Cortical connections of visual area MT in the macaque. *The Journal of Comparative Neurology, 248*, 190–222.

Ungerleider, L. & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle, M. Goodale & R. Mansfield (Eds.), *Analysis of Visual Behavior* (pp. 549–586). Cambridge: MIT Press.

Warrington, E. & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*, 829–854.

Wehner, R. (1992). Homing in arthropods. In F. Papi (Ed.), *Animal Homing* (pp. 45–144). London: Chapman and Hall.

Winston, P. (1975). Learning structural descriptions from examples. In P. Winston (Ed.), *The Psychology of Computer Vision.* New York: McGraw-Hill.

Zeki, S. (1993). *A Vision of the Brain*. Blackwell Scientific Publications.